Iranian Journal of Animal Science Research
Vol. 13, No. 3, Fall 2021, p. 453-462

نشریه پژوهشهای علوم دامی ایران
جلد ۱۳، شماره ۳، پاییز ۱۴۰۰، ص. ۴۶۲-۴۵۳

مقاله علمی - پژوهشی

مقایسه روش‌های تجزیه مؤلفه‌های اصلی (PCA) و تجزیه تشخیصی مؤلفه‌های اصلی (DAPC) در بررسی ساختار جمعیتی نژادهای اسب آخال‌تکه، عرب و کاسپین با استفاده از اطلاعات ژنومی

نسرین بابائی[۱]، سید عباس رافت[۲]، محمدحسین مرادی*[۳]، محمدرضا فیضی درخشی[۴]

تاریخ دریافت: ۱۳۹۸/۰۹/۲۴
تاریخ پذیرش: ۱۳۹۹/۱۰/۰۷

بابائی، ن.، س. ع. رافت، م. ح. مرادی، و م. ر. فیضی درخشی. ۱۴۰۰. مقایسه روش‌های تجزیه مؤلفه‌های اصلی (PCA) و تجزیه تشخیصی مؤلفه‌های اصلی (DAPC) در بررسی ساختار جمعیتی نژادهای اسب آخال‌تکه، عرب و کاسپین با استفاده از اطلاعات ژنومی. پژوهشهای علوم دامی ایران ۱۳(۳): ۴۶۲-۴۵۳.

چکیده

ابداع روش‌های تعیین ژنوتیپ با توان بالا و مقرون به صرفه طی سالیان اخیر، امکان ارزیابی ساختار ژنتیکی و ارتباط میان جمعیت‌های یک گونه را با استفاده از اطلاعات ژنومی فراهم ساخته است. بررسی ساختار جمعیتی بر اساس نشانگرهای گسترده در سطح ژنوم، اطلاعات ارزشمندی در ارتباط با روابط تکاملی و دسته‌بندی زیرجمعیت‌ها فراهم می‌کند. هدف از این تحقیق مقایسه دو روش تجزیه مؤلفه‌های اصلی (PCA) و تجزیه تشخیصی مؤلفه‌های اصلی (DAPC) در بررسی ساختار و روابط بین جمعیتی سه نژاد اسب موجود در منطقه خاورمیانه شامل آخال‌تکه، عرب و کاسپین بود. به این منظور از داده‌های ژنومی بدست آمده از آرایه‌های Illumina 50K SNP Beadchip در ۶۱ نمونه از این نژادها استفاده شد. این تحقیق با همکاری پروژه کنسرسیوم تنوع ژنتیکی اسب (EGDC) انجام شد و کدهای مورد نیاز برای آنالیز داده‌ها در نرم‌افزار R تهیه شدند. نتایج حاصل نشان داد که در هر دو روش ۱۰/۸ درصد واریانس توسط دو مؤلفه اول توجیه می‌شود و هر دو روش سه جمعیت را جدا از هم خوشه‌بندی کردند. معیار ارزیابی تعداد بهینه خوشه‌بندی برای روش DAPC، معیار اطلاعات بیزی (BIC) بود که تعداد K=3 بهترین نتیجه را با کمترین BIC نشان داد. روش DAPC نسبت به روش PCA با نتایج بهتری همراه بود. در تعیین شمار بهینهٔ K بهتر از روش PCA عمل کرد و تصویر بهتری از ارتباط بین افراد ارائه داد. همچنین در انتساب افراد به گروه‌های خود هر دو روش صحت بسیار خوبی ارائه دادند. در مجموع نتایج این تحقیق نشان می‌دهد با وجود اینکه نتایج تحقیقات گذشته این سه جمعیت را که مربوط به منطقه خاورمیانه هستند در یک خوشه از درخت هم‌سایگی قرار می‌دهند، ولی با توجه به نتایج این پژوهش و با استفاده از روش‌های مورد استفاده در این تحقیق، سه نژاد به صورت مجزا گروه‌بندی می شوند و DAPC می‌تواند تصویر بهتری از روابط بین جمعیتی در نژادهای اسب ارائه دهد.

واژه‌های کلیدی: روش‌های DAPC و PCA، ساختار جمعیت، نژادهای اسب خاورمیانه، نشانگرهای SNP.

مقدمه

حیوانات و گیاهان بومی به عنوان سرمایه ملی و ذخایر راهبردی هر کشور محسوب می‌شوند و حفظ و تکثیر آن‌ها از ارزش و اهمیت بسیاری برخوردار است. این موجودات پس از هزاران سال انتخاب طبیعی و مصنوعی و نیز گذر از موانع بسیار و با غلبه بر تمامی شرایط نامساعد محیطی همچنان به حیات

۱ و ۲- دانش آموخته کارشناسی ارشد ژنتیک و اصلاح نژاد دام و استاد گروه علوم دامی، دانشکده کشاورزی، دانشگاه تبریز، تبریز، ایران.
۳- دانشیار گروه علوم دامی، دانشکده کشاورزی و منابع طبیعی، دانشگاه اراک، اراک، ایران.
۴- دانشیار گروه کامپیوتر، دانشکده فنی و مهندسی، دانشگاه تبریز، تبریز، ایران.
(*- نویسنده مسئول: Email: Moradi.hosein@gmail.com)
DOI:10.22067/ijasr.0621.39343

خویش ادامه داده و به تکثیر و ازدیاد نسل پرداخته‌اند و همچنین نسبت به بسیاری از محدودیت‌های محیطی سازگاری پیدا کرده اند (۱ و ۱۱). حفاظت از ذخایر ژنتیکی بومی و مدیریت مؤثر این منابع ژنتیکی در دام‌های اهلی مبتنی بر مطالعه تنوع ژنتیکی توده‌های بومی (۹) و شناخت ساختار ژنتیکی جمعیت‌ها (۸) است. تعریف ساختار جمعیتی یک گونه برای حفاظت از یکپارچگی و هویت نژادی یک گونه مهم است و یک پیش‌شرط برای مدیریت منابع ژنتیکی محسوب می شود (۲۴ و ۲۹). استنتاج ساختار جمعیتی، در مدیریت مؤثر منابع ژنتیکی در نژادهای بومی (۸) و به دست آوردن دیدگاهی دربارهٔ تاریخچه و سیر تکامل جوامع ضروری است. علاوه بر این معین کردن ساختارهای جمعیتی و اختصاص افراد به جمعیت‌های مربوط به خود در مطالعات پویش پیوستگی ژنومی (Genome wide association study) می‌تواند در کاهش ارتباطهای کاذب مؤثر باشد (۲۲).

اسب یکی از کهن‌ترین گونه‌هایی است که در دشت‌های اوراسیا در حدود ۶۰۰۰–۵۰۰۰ سال قبل اهلی شده است (۱۷) و از نظر پراکنش جغرافیایی تقریباً در تمام سرزمین‌هایی که انسان در آن گام نهاده است این حیوان نیز در کنار او حضور داشته است. اسب دارای پراکنش جهانی است ولی با توجه به شرایط جغرافیایی خاص هر منطقه و کشور، نژادهای مقاوم به شرایط هر منطقه نیز شکل گرفته است. بر اساس آمارها بیش از ۵۰۰ نژاد برای این گونه دامی در سراسر جهان گزارش شده است (۱۷). شناخت ساختار ژنتیکی و روابط موجود میان نژادهای مختلف این گونه ارزشمند، می‌تواند در اولویت‌بندی برنامه‌های حفاظت ژنتیکی و اجرای برنامه‌های اصلاح‌نژادی دارای اهمیت باشد. در این تحقیق سه نژاد مهم اسب بومی آسیایی شامل آخال‌تکه، عرب و کاسپین مورد بررسی قرار گرفته است. این نژادها به ترتیب از کشورهای ترکمنستان، کشورهای عربی خاورمیانه و ایران منشأ می‌گیرند، هر چند در سایر کشورهای همجوار نیز پراکنش یافته‌اند.

امروزه اطلاعات حاصل از نشانگرهای SNP در سطح ژنوم به طور گسترده‌ای در مطالعه تنوع ژنتیکی دام‌های اهلی و ساختار جمعیتی استفاده می‌شود (۴ و ۲۷). داشتن ابزاری برای تجزیه و تحلیل ساختار جمعیتی با استفاده از این حجم اطلاعات نشانگری امری لازم است که روش‌های مختلفی برای بررسی ارتباط بین جوامع و قرار دادن افراد در جوامعی که از آن نشأت گرفته‌اند پیشنهاد شده است. روش‌های پرشماری برای تعیین ساختار ژنتیکی و لایه‌بندی جمعیت وجود دارد. یکی از شیوه‌های آماری برای آزمون ارتباط بین جمعیت‌ها و اختصاص افراد به آن‌ها با استفاده از ماتریس فاصله، استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA: Principal Component Analysis) است (۸) که قادر به تعیین ساختار جمعیتی می‌باشد (۱۵).

تجزیه و تحلیل مؤلفه‌های اصلی توسط پیرسون در سال ۱۹۰۱پیشنهاد شد (۱۶). پایهٔ کار این روش به این صورت است که در صفحه‌ی مختصات محورهای اصلی X و Y را طوری تغییر می‌دهد که مسیری در فضا پیدا شود تا مؤلفه‌های اصلی مربوط به داده‌ها در طول آن مسیر قرار گیرند. هر محوری که بزرگتر باشد نشان‌دهندهٔ آن است که واریانس بیشتری در میان داده‌ها در این جهت است و به همین دلیل آن‌را نخستین مؤلفه‌ی اصلی گویند. هدف از تجزیهٔ مؤلفه‌های اصلی آن است که واریانس موجود در داده‌های چندمتغیره را به مؤلفه‌هایی تجزیه کند که نخستین مؤلفه تا آنجا که ممکن است علت بیشترین واریانس موجود در داده‌ها باشد. دومین مؤلفه علت بیشترین واریانس ممکن پس از مؤلفه‌ی اول و الی آخر باشد (۱۶). تجزیه تشخیصی مؤلفه‌های اصلی (DAPC: Discriminant Analysis of Principal Components) روش چند متغیره است که برای شناسایی و توصیف خوشه‌های افرادی‌که به طور ژنتیکی با هم ارتباط دارند، طراحی شده است. این روش زمانی که گروه‌های پیشین وجود ندارند، k-means پی در پی و انتخاب مدل را برای استنباط ژنتیکی خوشه‌ها استفاده می‌کند. تجزیه تشخیصی مؤلفه‌های اصلی منجر به استخراج اطلاعات غنی از داده‌های ژنتیکی شده و انتساب دام‌ها به جمعیت را فراهم می‌کند. این رویکرد سریع‌تر از روش های خوشه‌بندی بیزی بوده و با طیف وسیعی از داده‌ها قابل اجراست (۵). یکی از مزایای DAPC سازگارپذیری بالای آن است. در واقع DAPC بر مدل ژنتیک جمعیت خاصی وابسته نیست و درنتیجه فرض تعادل هاردی واینبرگ و تعادل پیوستگی (لینکاژی) ندارد. به‌این ترتیب برای انواع موجودات صرف‌نظر از پلوئیدی بودن و نرخ نوترکیبی آن‌ها، کاربرد دارد (۵ و ۶).

هدف این تحقیق بررسی ساختار جمعیتی اسب‌های آخال‌تکه، عرب و کاسپین موجود در منطقه خاورمیانه با استفاده از داده‌های Illumina 50K SNP Beadchip و جداسازی این جمعیت‌ها با روش‌های تجزیه مؤلفه اصلی (PCA) و تجزیه تشخیصی مؤلفه های اصلی (DAPC) بود. این تحقیق با همکاری پروژه کنسرسیوم تنوع ژنتیکی اسب (Equine Genetic Diversity Consortium) انجام شده است. با توجه به پراکنش گسترده سه اسب آخال‌تکه (ترکمن)، عرب و کاسپین در ایران نتایج این تحقیق می‌تواند اطلاعات ارزشمندی در زمینه درک بهتر و مدل کردن معماری ژنتیکی نژادهای اسب ایرانی نیز فراهم آورد.

مواد و روش‌ها

نمونه‌گیری و تعیین ژنوتیپ:

در این مطالعه جهت بررسی ساختار جمعیتی برخی از نژادهای اسب آسیایی از اطلاعات ژنومی مجموع ۶۱ حیوان مربوط به نژادهای آخال‌تکه (۱۹رأس)، عرب (۲۴ رأس) و کاسپین (۱۸رأس) استفاده شد. داده‌های این تحقیق از پروژه کنسرسیوم تنوع ژنتیکی اسب

(Equine Genetic Diversity Consortium) تهیه شـد که نحوه جمع‌آوری و تعین ژنوتیپ نمونه‌ها در مقاله (۱۷) ارائه شده است، اما به طور خلاصه نمونه‌های مو یا بافت از حیوانات جمع‌آوری شدند. استخراج DNA با استفاده از روش بهینه‌یافته Pure gene (Qiagen) از بافت انجام شد و حدود ۱ میکروگرم DNA برای تعیین ژنوتیپ نمونه‌ها استفاده شد. تعیین ژنوتیپ نیز با استفاده از آرایه‌های Illumina SNP50KBead Chip که امکان تعیین ژنوتیپ ۵۲۶۰۳ جایگاه نشانگری SNP را فراهم می‌کند، بر اساس دستورالعمل استاندارد ایلومینا انجام شد.

ویرایش داده‌ها:

در این تحقیق برای اطمینان از کیفیت داده‌های حاصل از تعیین ژنوتیپ در آنالیزهای نهایی مراحل مختلف کنترل کیفیت بر روی داده های اولیه اعمال شد. ابتدا حیوانات با بیش از ۵٪ ژنوتیپ از دست رفته از آنالیزهای بعدی کنار گذاشته شدند و سپس SNPهایی که در مجموع حیوانات دارای فراوانی آلل نادر و نرخ خوانش (درصدی از نمونه‌ها که برای آن نشانگر تعیین ژنوتیپ شده‌اند) به ترتیب کمتر از ۲ درصد و ۹۵ درصد بودند (۱۰)، حذف شدند. در نهایت برای SNP های باقی مانده، آن هایی که خارج از تعادل هاردی واینبرگ بودند، به عنوان معیاری از خطای تعیین ژنوتیپ کنار گذاشته شدند (۲۶). برای بد ست آوردن سطح معنی‌داری در این آزمون از تصحیح بنفرونی ($\beta=\frac{\alpha}{n}$) استفاده شد. که در این فرمول n تعداد آزمون و همان تعداد نشانگرهای SNP (n=۵۰۰۰۰) و خطای نوع اول α=۰/۰۵ در نظر گرفته شدند که درنتیجه سطح احتمال معنی‌داری $\beta=10^{-6}$ خواهد شد. باید توجه کرد که هرچند عدم تعادل هاردی واینبرگ در جایگاه‌های مختلف الزاماً نشان‌دهنده خطای تعیین ژنوتیپ نیست و عوامل مختلفی همچون انتخاب، رانش و جهش نیز می‌توانند باعث انحراف از تعادل شوند، با این وجود همانطور که در مقاله Teo *et al.*, (۲۶) به تفسیر بیان شده است انحراف‌های شدید از این تعادل به احتمال زیاد به دلیل خطای تعیین ژنوتیپ و اشتباه در مراحل اتصال پرایمرها در هنگام ژنوتیپ نمونه‌ها با استفاده از آرایه‌های SNP Chip اتفاق می‌افتد (۱۰). در نهایت اطلاعات ژنوتیپی جهت وارد شدن به مراحل بعدی آنالیز، به فرمت ۰، ۱و ۲ (سه دسته ژنوتیپ) با نرم‌افزار plink (۱۹)تبدیل شدند.

آنالیزهای آماری

تجزیه مؤلفه‌های اصلی (PCA):

هدف از تجزیه به مؤلفه‌های اصلی آن است که واریانس موجود در داده‌های چند متغیره را به مؤلفه‌هایی تجزیه کند که اولین مؤلفه اصلی تا آنجا که ممکن است علت بیشترین واریانس موجود در داده‌ها را توجیه کند. دومین مؤلفه نیز به ترتیب بیشترین واریانس ممکن بعد از مؤلفه اول و الی آخر را توجیه نماید. به‌علاوه در این روش هر مؤلفه مستقل از مؤلفه‌های دیگر است، یعنی بین هر مؤلفه و مؤلفه‌های دیگر همبستگی وجود ندارد. به این منظور، ابتدا ماتریس واریانس-کواریانس متغیرهای مستقل محاسبه می شود. سپس بردار و مقدار ویژه این ماتریس محاسبه شده و بر اساس مقدارهای ویژه با اندازه بزرگ، بردار ویژه مربوطه به آنها انتخاب و به عنوان ضرایب ترکیب خطی متغییر اصلی جهت تشکیل توابع مؤلفه اصلی استفاده می شود بطوریکه این توابع از یکدیگر مستقل باشند. هرکدام از این توابع بخشی از واریانس متغییرهای مستقل را توجیه می‌کنند که خود تعیین‌کننده تعداد توابع مورد نیاز جهت حفظ حداکثری اطلاعات موجود در متغییرهای اصلی و همچنین میزان کاهش ابعاد داده‌ها می‌باشد (۳) در نهایت، با استفاده از این توابع مقدار آنها برای هر یک از مشاهدات محاسبه و از این داده‌های جدید برای تعیین و ترسیم الگو داده‌ها استفاده می‌شود. مراحل مختلف تجزیه‌ی PCA در این تحقیق، در محیط R و با به کارگیری تابع Princomp انجام شد.

بررسی ساختار جمعیت با استفاده از روش تجزیه‌ی چند متغیره تشخیص مؤلفه‌های اصلی (DAPC):

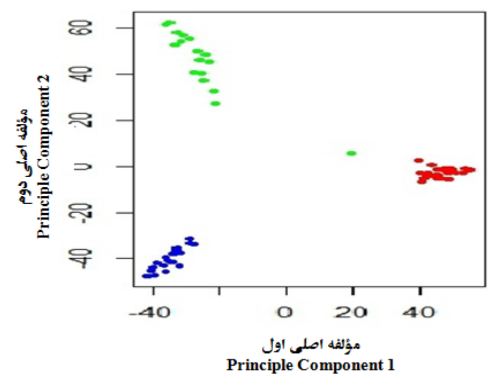
به‌منظور تشخیص تمایز و تنوع ژنتیکی در بین و داخل گروه‌ها، تجزیه‌ی تشخیص مؤلفه‌های اصلی (DAPC) که یک روش پارامتری و غیر مبتنی بر مدل می‌باشد، با بسته آماری adegenet مربوط به نرم‌افزار R اجرا شد. شناسایی تعداد بهینه خوشه با اجرای پی‌درپی و با افزایش K و راه‌حل‌های مختلف خوشه‌بندی در مقایسه با معیار اطلاعات بیزی (BIC: Bayesian Information Criterion) اجرا شد. در عمل بهترین BIC توسط نقطه حداقل در منحنی ارزش BIC به عنوان تابعی از K نشان داده شد. این تابع در آغاز داده‌ها را با PCA تبدیل می‌کند، سپس الگوریتم K-means با افزایش مقادیر K اجرا می‌شود (۵).

نتایج و بحث

مراحل مختلف ویرایش داده‌ها بر روی ۵۲۶۰۳ جایگاه نشانگری SNP در ۶۱ نمونه از نژادهای مختلف اسب اجرا شد. بعد از کنترل کیفیت اولیه، هیچ نمونه‌ای در جریان تعیین ژنوتیپ با بیش از ۵ درصد داده گم شده حذف نشد، مجموع تعداد ۱۱۸۳۰ نشانگر SNP به دلیل فراوانی آللی نادر و سه SNP به دلیل انحراف از تعادل هاردی-وینبرگ کنار گذاشته شدند. در نهایت، مجموع ۶۱ حیوان با ۴۰۷۷۰ نشانگر SNP مراحل مختلف کنترل کیفیت را با موفقیت گذرانده و برای مراحل بعدی آنالیز داده‌ها مورد استفاده قرار گرفتند.

تجزیه و تحلیل مؤلفه‌ی اصلی (PCA)

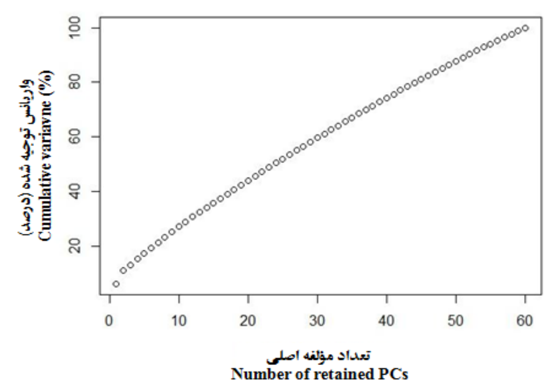
برای بررسی چگونگی تفکیک نژادها و ارزیابی اختلاف ژنتیکی میان جمعیت‌ها گراف مؤلفه‌های اصلی برای سه جمعیت مورد مطالعه در این تحقیق ترسیم شد (شکل ۱). همانطور که در شکل ۱ مشاهده می‌شود سه جمعیت در سه خوشه کاملا متمایز دسته‌بندی می‌شوند. نتایج تجزیه PCA بر اساس PC1 و PC2 نشان داد که این ۳ جمعیت (نژاد) هیچ هم‌پوشانی باهم ندارند.



شکل ۱– تجزیه PCA مربوط به سه نژاد اسب مورد مطالعه در این تحقیق بر اساس دو مؤلفه اصلی اول: افراد با نقطه‌های آبی، قرمز و سبز رنگ به ترتیب نشان‌دهنده نژادهای آخال‌تکه، عرب و کاسپین هستند.

Figure 1- PCA analysis based on the first two principal components in the three horse breeds investigated in this study: individuals with blue, yellow and red dot colors are showing Akhal-Take, Arabian and Caspian breeds, respectively.

دو مؤلفه اول ۱۰/۸ درصد واریانس را توجیه می‌کنند. ۵۳ مؤلفه‌ی اول ۹۰/۳ درصد واریانس را در این جمعیت‌ها توجیه می‌کند که بالا بودن واریانس توجیهی نشان‌دهنده‌ی این است که این جمعیت‌ها متمایز از هم هستند. برای توجیه ۱۰۰ در صد واریانس ۶۰ مؤلفه اول مورد نیاز است (شکل ۲).



شکل۲– واریانس تجمعی توصیفی توسط مؤلفه‌های اصلی: محور X تعداد مؤلفه اصلی و محور Y میزان واریانس توجیه شده را نشان می‌دهد.

Figure 2- Cumulative variance explained by different number of principal components: The X axis is showing the number of principal components and the Y axis is standing for the variance explained.

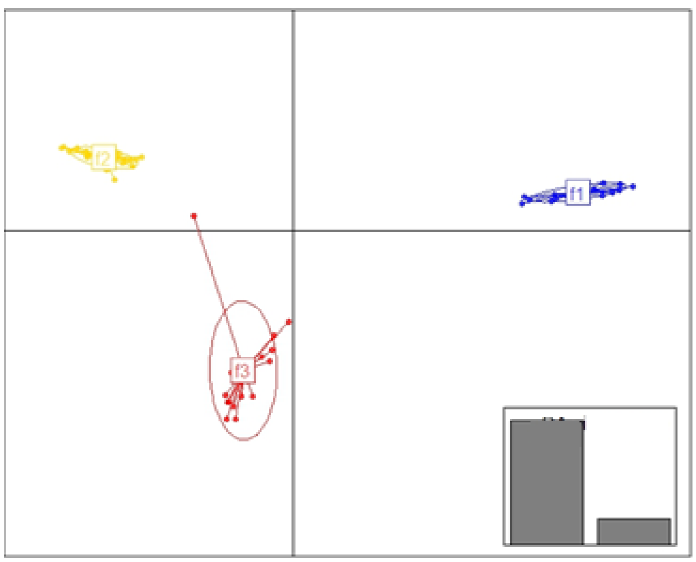
در بررسی تنوع ژنتیکی در نژاد های مختلف اسب دنیا با داده‌های SNP در سطح ژنوم، پتر سون و همکاران (۱۷) نشان دادند که نژادهای خاورمیانه، آخال‌تکه، عرب و کاسپین در یک محدوده از درخت همسایگی (Neighbor Joining) قرار می‌گیرند و همچنین گزارش شد که اسب‌های مغولی در گسترش این نژادها در سراسر آسیا تاثیرگذار بوده‌اند، درحالیکه در تحقیق حا ضر این سه جمعیت با توجه به اینکه از سه منطقه جغرافیایی متفاوت منشأ گرفته‌اند در سه جمعیت مجزا خوشه‌بندی می‌شوند. این نتیجه از روش PCA همراستا با نتایج محققانی همچون پاچو و همکاران (۱۴)، نُومبر و همکاران (۱۳) و ریچ و همکاران (۲۱) بود. آن‌ها بیان داشتند که روش PCA قادر به

شناسایی مهاجرت‌ها و جمعیت‌های ایزوله شده بوسیله عوامل مختلف هستند. روش PCA به عنوان جایگزینی برای روش‌های خوشه‌بندی بیزی نیز پیشنهاد شده است (۸).

پنج نشانگر SNP مهم که در مؤلفه اول بیشترین اهمیت را داشتند SNPهای chr1.40927347، chr1.36821928، chr1.13788925، chr1.30146112 و chr1.12056924 با ضرایب به ترتیب ۰/۰۱۰۷۹، ۰/۰۱۱۲۹، ۰/۰۱۱۴۳، ۰/۰۱۲۳ و ۰/۰۱۴۵۷ بودند و همچنین پنج نشانگر SNP نیز که کمترین اثر را در مؤلفه اول داشتند به ترتیب SNPهای chr1.41719058، chr126309054، chr1.41492293، chr1.14466144 و chr1.40502663 با ضرایب ۰/۰۱۰۸۱- ، ۰/۰۱۰۴-، ۰/۰۱۰۳- ، ۰/۰۱۰۲- و ۰/۰۰۹۵- بود ند. هرچند تعداد نمونه‌های اسب به کار گرفته شده در این مطالعه به دلیل بالا بودن هزینه‌های تعیین ژنوتیپ و تأمین داده‌ها از پروژه کنسرسیوم تنوع ژنتیکی اسب ممکن است نسبتا کم به نظر بیاید، اما باید دقت داشت که نتایج تحقیقات قبلی نشان می‌دهد که این تعداد نمونه برای انجام آنالیزهای مرتبط با ساختار ژنتیکی جمعیت‌ها کافی می‌باشد. گزارشات قبلی ثابت کرده‌اند که اندازه کوچک نمونه تنها می‌تواند موجب ایجاد اریب کوچکی در برآورد معیارهایی نظیر ساختار ژنتیکی جمعیت‌ها و تفرق جمعیتی شود، اما میزان این اریب در مورد معیاری مانند تنوع آللی بسیار بیشتر است (۲۵، ۲۸) به همین خاطر در این تحقیق از این داده‌ها در بررسی ساختار ژنتیکی نژادهای اسب خاورمیانه استفاده شد.

تجزیه‌ی تشخیصی مؤلفه‌های اصلی (DAPC)

شکل ۳ چگونگی خوشه‌بندی جمعیت‌های مختلف اسب مورد مطالعه در این تحقیق را با استفاده از روش DAPC با استفاده از دو مولفه تشخیصی اول نشان می‌دهد. نتایج حاصل نشان داد که دو تابع تشخیصی اول حدود ۱۰/۸ درصد واریانس را توجیه می‌کند.



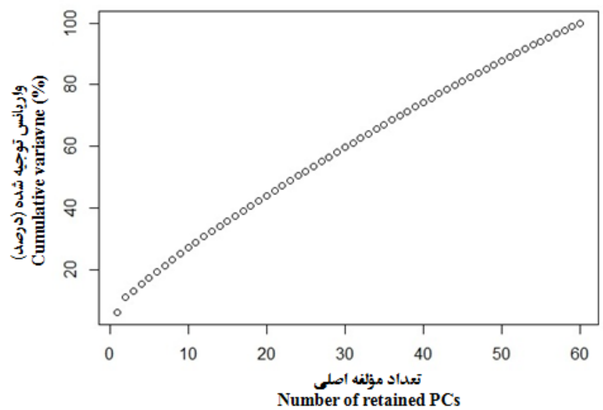
شکل۳- تجزیه و تحلیل مؤلفه‌های اصلی تشخیصی مربوط به نژادهای اسب مورد مطالعه در این تحقیق: خوشه‌ها با رنگ‌های مختلف مشخص شده‌اند و افراد با نقطه‌های آبی، زرد و قرمز رنگ به ترتیب نشان‌دهنده نژادهای آخال‌تکه، عرب و کاسپین هستند.

Figure 3- DAPC analysis for three breeds of horse evaluated in this study: Clusters are marked with different colors and individuals with blue, yellow and red dot colors are showing Akhal-Take, Arabian and Caspian breeds, respectively.

هر دو روش PCA و DAPC در این تحقیق باعث تفکیک صحیح جمعیتی در نژادهای اسب مورد مطالعه شدند. نابیلو و همکاران (۱۲) نیز در مطالعه‌ی ساختار جمعیتی گوسفندان ایرانی نتایج مشابهی را گزارش کردند. با این وجود، روش DAPC تداوم بالایی در تخمین ساختار جمعیتی و استنباط احتمالات عضویت افراد به هر گروه نشان داده است (۱۸). از آنجاکه روش PCA فاقد برخی از ویژگی‌های ضروری برای بررسی ساختار جمعیت‌های بیولوژیکی از جمله ناتوانی در ارزیابی گروهی است، لذا از آنجا که در اکثر تحقیقات در نظر است برای پی بردن به ساختار جمعیت با تعیین تعداد خوشه‌ها (گروه) بدون دانش قبلی اقدام شود، در این تحقیق، برای بررسی زیر جمعیت‌ها جدا از تاثیر مکان جغرافیایی از روش تجزیه تشخیصی مؤلفه اصلی (DAPC) نیز استفاده شد. در این روش محقق نظری در خصوص تعداد زیرجمعیت‌ها نمی‌دهد بلکه این روش آماری است که تعداد زیر جمعیت‌ها را پیشنهاد می‌دهد. رویکرد روش‌های ساختاری بر این

فرض استوارند که نشانگرها با هم مرتبط نبوده و جمعیت‌ها دارای آمیزش تصادفی هستند (۶). لذا استفاده از روش مستقل از مدل همانند DAPC روش راحت‌تری برای جمعیت‌های متمرکز محسوب می‌شوند. همچنین در تصحیح لایه‌بندی جمعیتی، روش DAPC به علت اینکه واریانس بین گروهی را افزایش و واریانس داخل گروهی را کاهش می‌دهد بهتر از روش PCA عمل می‌کند. نرخ انتساب صحیح یا صحت روش DAPC در انتساب افراد به گروه‌ها از ۸۰ در صد تا ۹۷ درصد بسته به تعداد تکرار متفاوت بوده است (۶). از سوی دیگر، این روش موجب کاهش زمان محاسباتی مورد نیاز جهت استخراج اطلاعات از داده‌های حجیم می‌شود (۷).

شکل ۳ حاکی از این است که روش DAPC قادر است تا تمام حیوانات را در سه گروه متمایز خوشه‌بندی کند. همچنین با توجه به خوشه‌هایی که تعیین می‌کند و مرکز هر خوشه را مشخص می‌کند تصویر واضح‌تری از گروه‌های دام‌ها نسبت به روش PCA که پیش‌تر اشاره گردید، نمایان می‌سازد. سزارامن (۲۳) نیز نشان داد که در استنباط و تفسیر ساختار ژنتیکی از روش‌های پرشمار بررسی ساختار جمعیتی، روش DAPC با توجه به برتری‌هایی همچون تفسیر بهتر زیر جمعیت‌ها و انتساب خوشه‌ها بهتر از روش PCA عمل می‌کند. تابع find. Cluster شکلی از واریانس تجمعی توصیف شده توسط مقدارهای ویژه مؤلفه‌های اصلی DAPC را توصیف می‌کند. نتایج نشان داد که ۶۰ مؤلفه اول صد در صد واریانس را توجیه می‌کنند (شکل ۴).



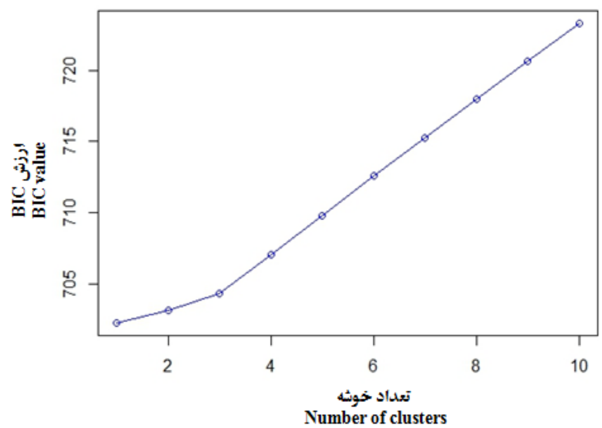
شکل٤- واریانس تجمعی توصیفی توسط مؤلفه‌های اصلی در روش DAPC: محور X تعداد مؤلفه اصلی و محور Y میزان واریانس توجیه شده را نشان می‌دهد.

Figure 4- Cumulative variance explained by different number of principal components for DAPC method: The X axis is showing the number of principal components and the Y axis is standing for the variance explained.

شکل ۵ مقادیر BIC را با افزایش مقادیر k نشان می‌دهد، طبق این شکل، کمترین مقدار BIC برای k=3 بدست آمد که بعد از این تعداد، BIC افزایش می‌یابد و به طور واضح نشان می‌دهد که این تعداد خوشه بهتر است در نظر گرفته شود. بعد از K=3 کمترین مقدار BIC برای خوشه ۱۴ است و رفته رفته این مقدار افزایش پیدا می‌کند. این معیار با توجه به شکل‌های بالا وجود یک خوشه را تایید می‌کند و اگر بخواهیم به چند خوشه تقسیم‌بندی کنیم نهایت سه خوشه با توجه به شکل با BIC بالاتر از K=3 قابل تفکیک هستند. علاوه براین، DAPC احتمالات اعضای هر فرد را برای گروه‌های مختلف براساس توابع تشخیصی جایگشتی فراهم می‌آورد و می‌توانند نزدیکی دام‌ها به خوشه‌های مختلف را تفسیر کنند (۵). احتمالات عضویت دام‌ها، خوشه های ژنتیکی صریح و واضحی را فراهم می‌آورد. کریمی (۷) با هدف بررسی ساختار ژنتیکی جمعیت گاوهای بومی ایران بر اساس اطلاعات ژنومی، تعداد ۹۰ رأس گاو از هشت جمعیت مختلف از گاوهای بومی کشور شامل نژادهای سرابی، سیستانی، کردی، کرمانی، مازندرانی، تالشی، نجدی و پارس را با استفاده از روش DAPC مورد بررسی قرار داد. بر اساس نتایج حاصل تعداد سه خوشه ژنتیکی بر اساس روش K- means در بین گاوهای بومی ایران نشان داده شد. جمعیتهای مازندرانی و تالشی در انطباق با توزیع جغرافیایی این نژادها، در خوشه یکسانی قرار گرفتند. همچنین، نژادهای پراکنده شده در مناطق کوهستانی شمال غرب کشور (سرابی و کردی) و نژادهای مناطق جنوبی کشور (سیستانی، کرمانی، پارس و نجدی) دو گروه ژنتیکی متمایز دیگر را تشکیل دادند. بر اساس نتایج این تحقیق روش DAPC روشی مناسب برای بررسی ساختار جمعیتی نژادهای بومی گزارش شده است.

همانطور که قبلا اشاره شد پترسون و همکاران (۱۷) در بررسی تنوع ژنتیکی نژادهای اسب دنیا گزارش کردند که نژادهای آخال‌تکه، عرب و کاسپین در یک خوشه قرار می‌گیرند و بیان شد که جمعیت اسب مغولی و توا در گسترش اسب‌ها در سراسر آسیا و اروپا تاثیرگذار

هستند که این نتایج می‌تواند ناشی از سطح بالای تنوع درون نژادی در این جمعیت‌ها باشد. لذا برای ارزیابی ارتباط بین خوشه‌های مختلف، باید یک روش مناسب بر تنوع بین گروه‌ها تمرکز کند و تنوع درون گروهی را نادیده بگیرد که تجزیه‌ی تشخیصی مؤلفه‌های اصلی (DAPC) این کار را انجام می‌دهد. در مطالعه بر روی گاومیش‌های ایران روش‌های تجزیه وتحلیل مؤلفه‌های (PCA) و تجزیه وتحلیل تشخیصی مؤلفه‌های اصلی (DAPC) اجرا شد. روش DAPC توانست با معیار ارزیابی، BIC K=3 بهترین نتیجه را نشان دهد. در این تحقیق روش DAPC در بررسی ساختار جمعیتی نسبت به روش PCA بهتر گزارش شده است و در کنترل کیفیت و تصحیح لایه‌بندی جمعیتی در بررسی‌های پیوستگی به عنوان جایگزینی برای PCA پیشنهاد شده است (۲).



شکل ۵- تعداد خوشه‌ها برای جمعیت‌های اسب مورد مطالعه در این تحقیق بر اساس معیار اطلاعات بیزی: محور X تعداد خوشه و محور Y ارزش BIC را نشان می‌دهد.

Figure 5- Number of clusters for the horse populations investigated in this study based on Bayesian Information Criterion (BIC): The X axis is showing the number of clusters and the Y axis is standing for BIC values.

نتیجه‌گیری کلی

در این تحقیق دو روش تجزیه مؤلفه اصلی (PCA) و تجزیه تشخیصی مؤلفه اصلی (DAPC) برای تفکیک نژادی و شناسایی روابط بین جمعیتی نژادهای اسب آخال‌تکه، عرب و کاسپین مورد مقایسه قرار گرفتند که با توجه به پراکنش گسترده این سه نژاد در ایران نتایج این تحقیق می‌تواند اطلاعات ارزشمندی در زمینه درک بهتر ساختار ژنتیکی این نژادها فراهم آورد. این تحقیق با همکاری پروژه کنسرسیوم تنوع ژنتیکی اسب (EGDC) و با استفاده از اطلاعات حاصل از تراشه‌های ژنومی انجام شد. نتایج نشان داد که هرچند هر دو روش PCA و DAPC توانستند سه جمعیت را به سه نژاد خوشه‌بندی کنند و مشاهده شد که این سه نژاد کاملا از هم متمایز هستند اما در مجموع روش DAPC به دلیل به تصویر کشیدن تفاوت بین گروهی با زمان محاسباتی کمتر، و عدم نیاز به پیش‌فرض هایی دربارهٔ مدل ژنتیک جمعیت‌های مورد مطالعه، تصویر واضح‌تری از ساختار جمعیتی نژادها با استفاده از اطلاعات ژنومی ارائه می‌دهد و روش DAPC در تعیین تعداد بهینه K کارآمدتر عمل می‌کند.

تشکر و قدردانی

نویسندگان مقاله از همکاری صمیمانه پروژه کنسرسیوم تنوع ژنتیکی اسب (EGDC) و به خصوص خانم دکتر پترسون و همکاران ایشان به خاطر در اختیار قرار دادن اطلاعات ژنوتیپی نژادهای مورد مطالعه در این تحقیق کمال تشکر را دارند.

منابع

1. Askari, N., A. M. Mohammadi, and A. Baghizadeh. 2011. ISSR markers for assessing DNA polymorphism and genetic characterization of cattle, goat and sheep populations. Iranian Journal of Biotechnology, 9: 222-229.
2. Azizi, Z., H. Moradi-Shahrbabak, and M. Moradi-Shahrbabak. 2017. Comparison of PCA and DAPC methods for analysis of Iranian Buffalo population structure using SNPchip90k data. Iranian Journal of Animal Science, 2:153-161. (In Persian).
3. Dodds, K. G., B. Auvray, S. N. Newman, and J. C. McEwan. 2014. Genomic breed prediction in New Zealand sheep. BMC Genetics, 15:92.
4. Epps, C. W., J. A. Castillo, A. Schmidt-Küntzel, P. du Preez, G. Stuart-Hill, M. Jago, and R. Naidoo. 2013.

Contrasting historical and recent gene flow among African buffalo herds in the Caprivi Strip of Namibia. Journal of Heredity, 104(2): 172-181.
5. Jombart, T., and C. Collins. 2015. A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. Imp Coll London-MRC Cent Outbreak Anal Model, 43.
6. Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics, 11(1): 94.
7. Karimi, K. 2017. Investigation of the population structure in Iranian native cattle using discriminant analysis of principal components. Research on Animal Production, 8:184-193. (In Persian).
8. Liu, N., and H. Zhao. 2006. A non-parametric approach to population structure inference using multi locus genotypes. Human Genomics, 2(6): 353.
9. Mohammadi, A., M. R. Nassiry, J. Mosafer, M. R. Mohammadabadi, and G. E. Sulimova. 2009. Distribution of BoLA-DRB3 allelic frequencies and identification of a new allele in the Iranian cattle breed Sistani (Bos indicus). Russian Journal of Genetics, 45(2): 198-202.
10. Moradi, M. H., A. Nejati-Javaremi, M. Moradi-Shahrbabak, K. G. Dodds, and J. C. McEwan. 2012. Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC Genetics, 13: 10.
11. Moradi, M. H., J. Rostamzadeh, A. Rashidi, Kh. Vahabi, and H. Farahmand. 2013. Analysis of genetic diversity in Iranian Mohair goat and its color types using Inter Simple Sequence Repeat (ISSR) markers. Journal of Agricultural Communications, 1: 2.
12. Nabiloo, R., M. B. Zandi and M. T. Harakinezhad. 2018. Study of genetic diversity of indigenous (Afshari, Moghani and Ghezel) and exotic (Romney, Merinos and Dorper) sheep breeds using high-density SNP markers. Iranian Journal of Animal Science, 49(3):437-451. (In Persian).
13. Novembre, J., and M. Stephens. 2008. Interpreting principal component analyses of spatial population genetic variation. Nature Genetics, 40(5): 646-649.
14. Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-Correlated SNPs for structure identification in worldwide human populations. PLoS Genetics, 3(9): e160.
15. Patterson, N., A. L. Price, and D. Reich. 2006. Population structure and eigen analysis. PLoS Genetics, 2(12): e190.
16. Pearson, K. 1901. On lines and planes of closest fit to systems of point in space. Philosophical Magazine, 2: 559-572.
17. Petersen, J. L., J. R. Mickelson, E. G. Cothran, L. S. Andersson, J. Axelsson, E. Bailey, D. Bannasch, M. M. Binns, A. S. Borges, P. Brama, A. C. Machado, O. Distl, M. Felicetti, L. F. Clipsham, K. T. Graves, and et al. 2013. Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLOS ONE, 8(1): e54997.
18. Pometti, C. L., C. F. Bessega, B. O. Saidman, and J. C. Vilardi. 2014. Analysis of genetic population structure in Acacia caven (Leguminosae, Mimosoideae), comparing one exploratory and two Bayesian-model-based methods. Genetics and Molecular Biology, 37(1):64-72.
19. Purcell S. 2010. PLINK version 1.07. URL: http://pngu.mgh.harvard.edu/~purcell/pink.
20. Rahmaninia, J., S. R. Miraei-Ashtiani, H. Moradi-Shahrbabak. 2015. Unsupervised clustering analysis of population and subpopulation structure using dense SNP markers. Iranian Journal of Animal Science, 46(3): 277-278. (In Persian).
21. Reich, D., A. L. Price, and N. Patterson. 2008. Principal component analysis of genetic data. Nature Genetics, 40: 491.
22. Serre, D., A. Montpetit, G. Paré, J. C. Engert, S. Yusuf, B. Keavney, T. J. Hudson, and S. Anand. 2008. Correction of population stratification in large multi-ethnic association studies. PloS one, 3(1): e1382.
23. Sethuraman, A. 2013. On inferring and interpreting genetic population structure-applications to conservation, and the estimation of pairwise genetic relatedness. Ph.D. dissertation, Iowa State University, Iowa State.
24. Shojaei, M., M. Mohammad-Abadi, M. Asadi-Fozi, O. Dayani, A. Khezri, and M. Akhondi. 2011. Association of growth trait and leptin gene polymorphism in Kermani sheep. Journal of Cell and Molecular Research, 2(2): 67-73.
25. Smith, O., and J. Wang. 2014. When can noninvasive samples provide sufficient information in conservation genetics studies? Molecular Ecology Resources, 14: 1011-1023.
26. Teo, Y. Y., A. E. Fry, T. G. Clark, E. Tai, and M. Seielstad. 2007. On the usage of HWE for identifying genotyping errors. Annals of Human Genetics, 71(5): 701-703.
27. Uzzaman, M. R., Z. Edea, M. S. A. Bhuiyan, J. Walker, A. K. Bhuiyan, and K. S. Kim. 2014. Genome-wide single nucleotide polymorphism analyses reveal genetic diversity and structure of wild and domestic cattle in Bangladesh. Asian-Australasian Journal of Animal Sciences, 27(10): 1381.
28. Willing, E., C. Dreyer, and C. Van Oosterhout. 2012. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. PLoS One, 7(8): e42649.
29. Zamani, P., M. Akhondi, M. R. Mohammadabadi, A. A. Saki, A. Ershadi, M. H. Banabazi, and A. R. Abdolmohammadi. 2011. Genetic variation of Mehraban sheep using two inter simple sequence repeat (ISSR) markers. African Journal of Biotechnology, 10(10): 1812-1817.

Iranian Journal of Animal Science Research
Vol. 13, No. 3, Fall 2021, p. 453-462

نشریه پژوهشهای علوم دامی ایران
جلد ۱۳، شماره ۳، پاییز ۱۴۰۰، ص. ۴۶۲-۴۵۳

Comparison of principal component analysis (PCA) and discriminant analysis of principal component (DAPC) methods for analysis of population structure in Akhal-Take, Arabian and Caspian horse breeds using genomic data

Nasrin Babaei[1], Abbas Rafat[2], Mohammad Hossein Moradi*[3], Mohammad Reza Feizi Derakhshi[4]

Submitted: 15-12-2019
Accepted: 27-12-2020

Babaei, N., A. Rafat, M. H. Moradi, and M. R. Feizi Derakhshi. 2021. Comparison of principal component analysis (PCA) and discriminant analysis of principal component (DAPC) methods for analysis of population structure in Akhal-Take, Arabian and Caspian horse breeds using genomic data. Iranian Journal of Animal Science Research 13(3):453-462.

Introduction Development of high-power and cost-effective genotyping methods in recent years has provided the possibility of evaluation the genetic structure and the relationship among species populations utilizing genomic data. Genome wide inference of population structure using genetic markers could provide invaluable information associated with evolutionary relationships and clustering of subpopulations for performing animal breeding programs. In large scale studies, one of the interesting subjects is to study the existence of genetic differences among subdivided groups ascertained from different geographic locations. The objective of this study was to compare the principal component analysis (PCA) and discriminant analysis of principal component (DAPC) approaches for determining the population structure and study how an individual allocated to the true population of origin, in three Horse breeds located in Middle East consisting Akhal Take, Arabian and Caspian using genomic data.

Materials and Methods In this study, the genomic data obtained from 61 animals consisting Akhal Take (19), Arabian (24) and Caspian (18) were used to investigate the population structure of some Asian horse breeds. The data were obtained from the Equine Genetic Diversity Consortium (EGDC) project. Hair or tissue samples were collected from animals. DNA extraction was performed using an optimized Pure gene (Qiagen) assay and approximately 1 μg of DNA was used for genotyping of the samples. Genotyping was performed using Illumina SNP 50K BeadChip arrays that allow to genotype 52603 SNP marker loci, according to the Illumina standard guidelines. In this study, different quality control steps were applied on preliminary data to ensure the quality of genotyping data. Quality control carried out using PLINK v.1.07 program. The samples with more than 5% missing data were excluded from analysis. Then for each SNP, MAF and call percentage were calculated and the SNPs with a call rate<95% and a MAF<2% were discarded. Deviation from Hardy-Weinberg equilibrium ($p<10^{-6}$) was estimated for the remaining SNPs to identify genotyping errors. The Bonferroni correction ($\beta=\alpha/n$) was used to address the multiple testing comparison problem. Principal component analysis (PCA) is a statistical technique for summarizing data from many variables into a few variables which describe as much of the variation in the data as possible. For this purpose, the variance-covariance matrix of independent variables was first calculated and principal components were extracted. Each new variable has an associated Eigen value that measures the respective amount of explained variance. Furthermore, the model independent of discriminant analysis of principal component (DAPC) is a multivariate method designed to identify and describe clusters of genetically related individuals. When group priors are lacking, DAPC uses sequential K-means and model selection to infer genetic clusters. Analysis was performed using PCA and DAPC approaches and the codes for analysis were provided in

1- M.Sc. Graduate of Animal Science, Faculty of Agriculture, University of Tabriz, Tabriz, Iran.
2- Professor, Department of Animal Science, Faculty of Agriculture, University of Tabriz, Tabriz, Iran.
3- Associat Professor, Department of Animal Science, Faculty of Agriculture, Arak University, Arak, Iran.
4- Associat Professor, Department of Artificial intelligence, Faculty of Computer Engineering, University of Tabriz, Tabriz, Iran.
(*Corresponding author email: hoseinmoradi@ut.ac.ir)
DOI:10.22067/ijasr.0621.39343

R v.3.4.1 software.

Results and Discussion The analysis of the main components summarizes the general variation among individuals, which includes both the variability between the groups and the diversity of the groups, and shows a clear picture of the differences between the groups. The results of this study indicated that 10.8% of the variance was explained by the first two components in both PCA and DAPC methods. Both methods showed high accuracy for assigning of individuals to the true population of origin and both were able to cluster three populations separately. The Bayesian information criterion (BIC) index was used for evaluating the optimal number of clusters for DAPC method and the results revealed that K=3 showing the optimal number with lowest BIC that completely separate three populations. The DAPC method was better than PCA to separate populations from each other due to the increase of intergroup variance and the reduction of intra-group variance. In determining the optimal number of K, it worked better than PCA method and provided a better picture of the relationship between individuals. This results show that DAPC method can be applied in quality control of GWAS as an alternative to the PCA, because of summarizing the genetic differentiation between groups and overlooking within-group variation and provides better population structure.

Conclusion In general, the results of this study showed that although the previous studies grouped these three breeds located in Middle East in one cluster of neighboring trees, however, according to the results of this study, three breeds are grouped separately, and the DAPC method can better illustrate the inter-population relationships in horse breeds.

Keywords: Population structure, Middle East horse breeds, SNP markers, PCA and DAPC methods.